Gene Ontology-driven similarity for supporting the prediction of integrated functional networks

Francisco Azuaje

University of Ulster, UK

Goals of this research

- To propose a method to incorporate GO-driven information into the inference of functional networks
- To study their properties and relationships with other predictive resources
- To estimate its statistical and biological relevance

- **Our hypotheses:**
  - GO-driven similarity networks (GOSN) represent significant features of real functional networks
  - These networks, in combination with other relevant predictive resources, may improve the overall predictive ability of integrated networks

Rationale:
Post-genome biology (systems biology)

- Networks of functional relationships between genes and proteins based on different properties or resources, e.g. gene co-expression.
- A node in a network represents a gene. A connection is established if the nodes are significantly associated.
- Overlaps between different types of relationships support the idea of combining them to build more meaningful networks.
- For example, physically interacting proteins are more likely to have similar gene expression patterns, etc.

Rationale:
The role of functional annotations

- Functional annotations of gene products (e.g. annotations derived from GO-driven databases) have been recently proposed to support network inference.

- The application of GO-derived information to support the prediction of functional networks of genes has not been rigorously investigated.

- Comprehensive studies on the predictive properties of such networks have not been reported.

This remaining of this presentation

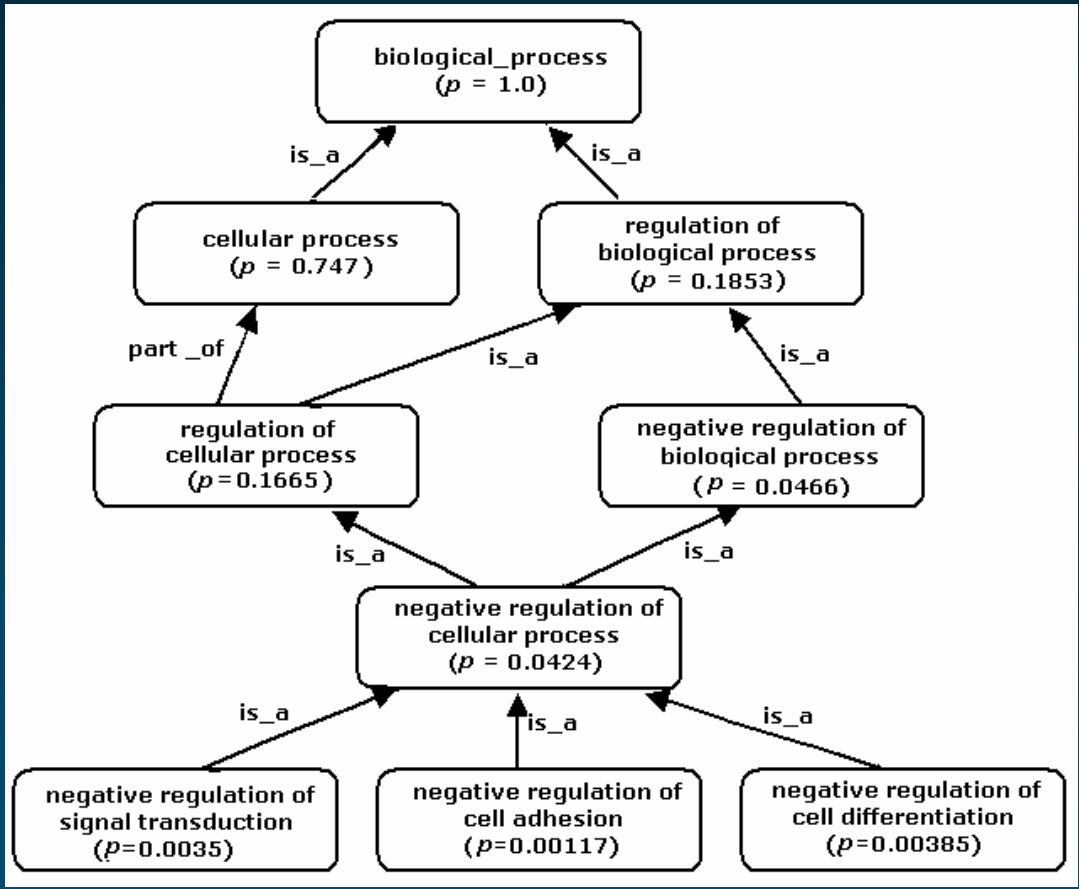
- Brief introduction to the Gene Ontology (GO) and its applications
- Estimating functional similarity with the GO
- Constructing GO-driven similarity networks (GOSN)
- Integrating GOSN and other single-source networks
- Some relevant results
- Current/future work and conclusions

The Gene Ontology

- Provides structured, controlled vocabularies that can be used to describe gene products in different organisms
- GO hierarchies: *Molecular function* (MF), *biological process* (BP), and *cellular component* (CC).
- MF: The role played by individual gene products, e.g. *G-protein coupled receptor activity*.
- BP: Objective accomplished by one or more ordered assemblies of molecular function, e.g. *signal transduction*.
- CC: Cellular localization of the gene product, e.g. *nucleus* or *anaphase-promoting complex*.

The GO

- GO terms and their relationships within each hierarchy form a network in which each term has one or more parent terms.
- The relationship between a child and its parent can be either “is a” (is a kind of) or “part of”.



Partial view of the GO Biological Process hierarchy

The GO and its applications

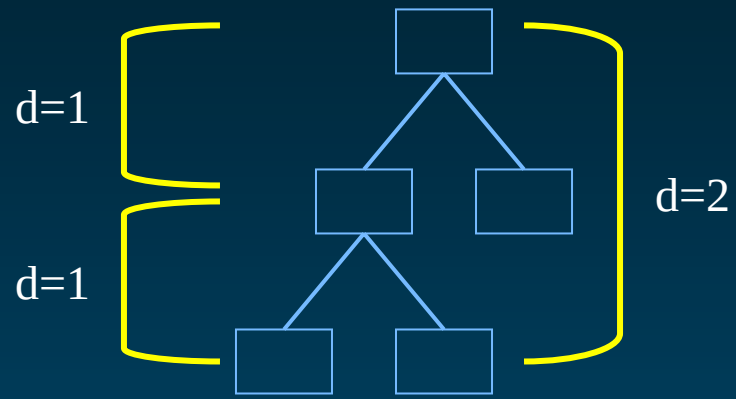
- Incorporation of GO annotations into gene expression data clustering analysis (significance of over-represented terms)
- Inference of gene-phenotype associations
- Assignment of new annotations to genes using gene expression and GO annotations
- Gold-standard in network prediction studies
- Predictive source for integrated network prediction

The GO and its applications (II)

- Estimating functional similarity using the GO and model organism databases annotated to GO (SGD, MGD, WB, etc.)
- Relationships between GO-driven similarity and sequence similarity, gene co-expression, functional interactions.
- We propose to build GOSN using non-traditional similarity assessment methods

Approaches to computing GO-driven similarity

- Edge counting
  - Intuitive
  - Requires density to be homogeneous in the taxonomy

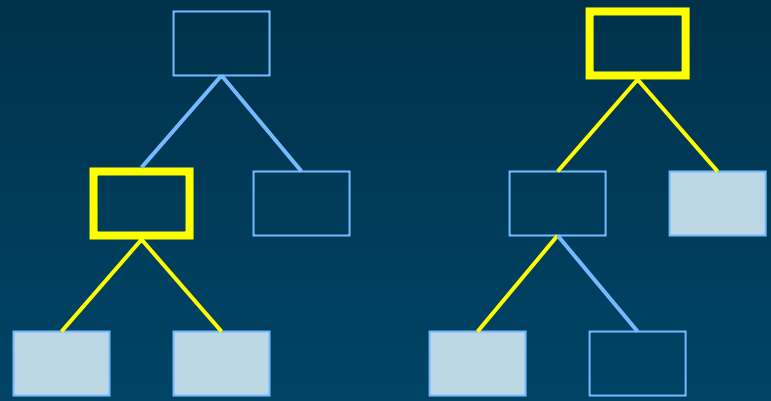


- Information-theoretic metrics
  - Grounded in information theory
  - Compensates for heterogeneity in the taxonomy

Slide courtesy of O. Bodenreider

Information-theoretic approaches

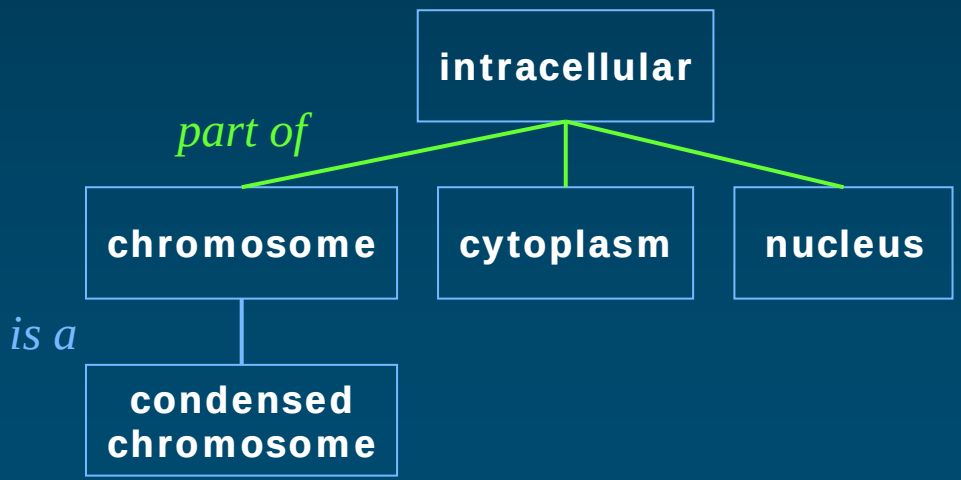
- *Information content (IC)*: nodes high in the hierarchy have a small IC
- The information shared by two nodes can also be represented by their common ancestors (*least common subsumer*)
- The more information two terms share, the more similar they are



Slide courtesy of O. Bodenreider

Information content in GO

- "Taxonomy":
  hierarchy (DAG) of *is a* + *part of* relations
- Frequency distribution of GO terms: annotation databases



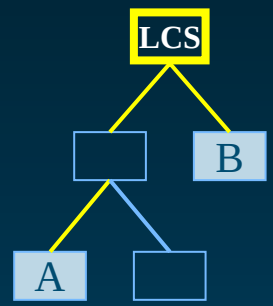
ZTA1	GO:0005634
ZTA1	GO:0005737
ZUO1	GO:0003754
ZUO1	GO:0005737
ZUO1	GO:0005829
ZUO1	GO:0005840
ZUO1	GO:0006457
ZWF1	GO:0004345
ZWF1	GO:0005737
ZWF1	GO:0006098

Slide courtesy of O. Bodenreider

GO-driven similarity

[Lord *et al.*, PSB 2003]
[Wang *et al.*, CIBCB 2004]

- Based on the information content of the least common subsumer (LCS)
- Several variants



- Resnik (1995)

$$sim(A,B) = \max_{LCS \in S(A,B)} \left[-\log p(LCS) \right]$$

- Lin (1998)

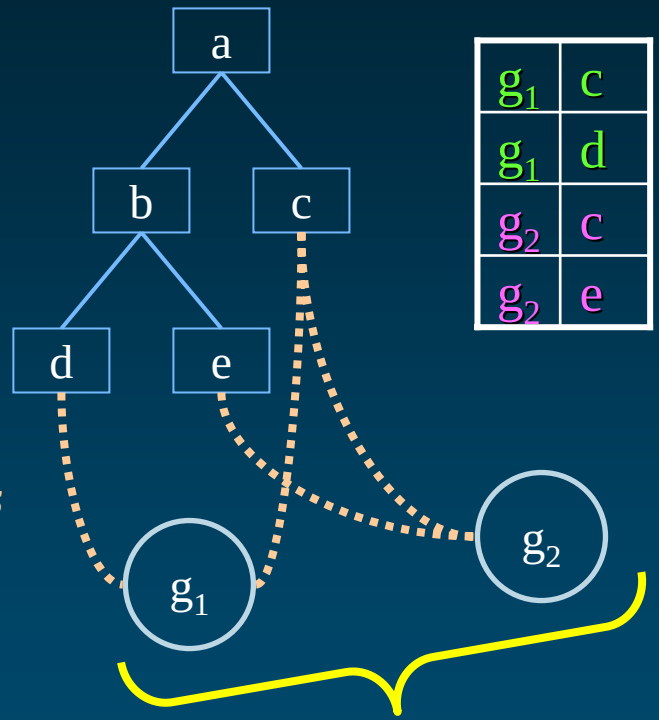
$$sim(A,B) = \frac{2 \times \log p(LCS)}{\log p(A) + \log p(B)}$$

- Jiang & Conrath (1997)

$$dist(A,B) = \log p(A) + \log p(B) - 2 \times \log p(LCS)$$

Slide courtesy of O. Bodenreider

GO-driven similarity among gene products



g_1	c
g_1	d
g_2	c
g_2	e

sim(c,c)
sim(c,e)
sim(d,c)
sim(d,e)

$SIM(g_1,g_2)$

$$SIM(g_i, g_j) = \frac{1}{m \times n} \times \sum_{c_k \in Ai, c_p \in Aj} sim(c_k, c_p)$$

Slide courtesy of O. Bodenreider

Constructing GOSN (I)

- GO annotations from the SGD
- Annotations encoded in the GO Biological Process hierarchy
- 57,367 pairs of genes with significant mRNA expression correlations originating from a comprehensive compendium of microarray data

Constructing GOSN (II)

- *Low similarity network (LSN)*: a connection between a pair of genes was established if their GOS was larger than 0 under the Biological Process hierarchy.
- *Medium similarity network (MSN)*: a connection between a pair of genes was established if their GOS was larger or equal to 0.5.
- *High similarity network (HSN)*: a connection between a pair of genes was established if their GOS was larger or equal to 0.8.
- *Very high similarity network (VHSN)*: a connection between a pair of genes was established if their GOS was equal to 1.

Constructing GOSN (II)

GOS networks vs. random networks (Mean similarity ± S.E)

GOS networks		Random networks	Sig.
LSN	0.374 ± 2.0E-03	0.150 ± 4.80E-04	$p < 0.001$
MSN	0.857 ± 2.0E-02	0.289 ± 1.0E-03	$p < 0.001$
HSN	0.98 ± 6.0E-4	0.48 ± 2.7E-03	$p < 0.001$
VHSN	1.0 ± 0.0	0.594 ± 2.0E-03	$p < 0.001$

S.E: Standard error, Sig.: Significance of the difference (Student's t test).

Other networks integrated (SSN)

- *SGA* network (genetic interactions) (Tong *et al*., 2004).
- *Homol* network: protein similarity (Altschul *et al*., 1997) (Zhang *et al*., 2005).
- *Coex* network: Highly co-expressed pairs of genes (Hughes *et al*., 2000).
- *Physic* network: pairs of proteins belonging to the same protein complex (Mewes *et al*., 2002; Gavin *et al*., 2002; Ho *et al*., 2002).
- *Chip* network: transcription factor-gene interactions (Tong *et al*., 2004).

Construction of integrated networks (I)

- Different integrated networks obtained by merging all types of single-source relationships (union of networks).

- Four networks were first obtained: *intLSN*, *intMSN*, *intHSN* and *intVHSN*, which were derived from the combination of the SSN with *LSN*, *MSN*, *HSN* and *VHSN* respectively.

Construction of integrated networks (II)

- Reference integrated network, *intNonGOS*, which did not incorporate the GOS networks.
- *Multiple-support integrated networks*, i.e. edges supported by at least two types of functional interactions; e.g. *intMSN-MS* is a multiple-support, integrated network that incorporates the *MSN*.

Detection of potential functional modules through network clustering

Clustering of networks: Summary description

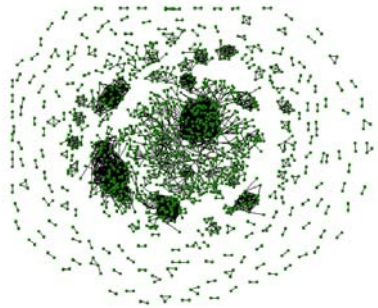
Network	NC	AC-score	AID	ANP	NC-score-5
MSN	51	3.83	91.56	12.41	9
HSN	36	3.85	84.64	10.97	8
VHSN	32	3.99	90.25	11.25	7
intLSN	-	-	-	-	-
intMSN	-	-	-	-	-
intHSN	-	-	-	-	-
intVHSN	-	-	-	-	-
intLSN-MS	-	-	-	-	-
intMSN-MS	53	3.96	99.26	15.54	11
intHSN-MS	51	3.39	75.18	13.16	9
intVHSN-MS	52	3.30	71.90	12.11	9
intNonGOS	-	-	-	-	-
intNonGOS-MS	38	2.92	41.23	10.21	5

NC: Number of clusters; AC-score: Average MCODE cluster score; AID: Average interaction density per cluster; ANP: Average number of proteins per cluster; NC-score-5: Number of clusters with MCODE cluster scores greater than 5.

Linking networks to significant functional categories and pathways (I)

intNonGOS-MS: Linking clusters to MIPS functional categories and KEGG pathways

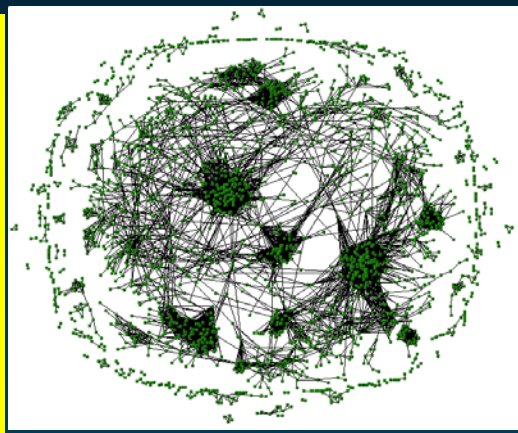
Cluster	Sample of significantly-represented MIPS functional categories (number of proteins)	Associations with KEGG pathways (number of proteins)
1	Stress response (9); Extracellular/secretion protein (1); Cell membrane (1);Unclassifierd (12)	-
2	Regulation of splicing (1);Ribosome biogenesis (25);ribosomal proteins (25);nucleic acid binding (6);RNA binding (6).	Ribosome (25)
3	Transposable elements, viral and plasmid proteins (19)	-
4	Transcription (9);rRNA processing (8);Ribosome biogenesis (13);ribosomal proteins (5);nucleic acid binding (6);RNA binding (6).	Ribosome (1)
5	DNA processing (6);DNA synthesis and replication (6);DNA topology (6);DNA recombination (6);Stress response (6);Biogenesis of nucleus (6);Organization of chromosome structure (6)	-



Linking networks to significant functional categories and pathways (II) (5 out of 11 clusters)

intMSN-MS: Linking clusters to MIPS functional categories and KEGG pathways

Cluster	Examples of significantly represented MIPS functional categories (number of proteins)	Associations with KEGG pathways (number of proteins)
1	Extracellular/secretion proteins (1); Cell membrane or cell wall attached (1);Unclassified proteins (32)	Galactose metabolism (2); Starch and sucrose metabolism (2);
2	rRNA processing(35);ribosome biogenesis (17); ribosomal proteins (8);nucleic acid binding (18); RNA binding (18);Nucleotide binding (7); ATP binding (7);	Ribosome (2)
3	Ribosome biogenesis (34); ribosomal proteins (34);nucleic acid binding (5);RNA binding (5).	Ribosome (36)
4	Amino acid metabolism (29);Assimilation of ammonia (6);Metabolism of glutamine (1);Degradation of glutamine (1);Metabolism of arginine (5);Biosynthesis of arginine (5); Metabolism of urea cycle (2);Metabolism of the aspartate family (9);Metabolism of threonine (3);Metabolism of methionine (4);Metabolism of serine (3);Metabolism of the pyruvate family (5);C-compound, carbohydrate anabolism (7); Secondary metabolism (7); Complex cofactor/cosubstrate binding (6)	Valine, leucine and isoleucine biosynthesis (4); Lysine biosynthesis (5); Phenylalaine, tyrosine and triptophan biosynthesis (8)
5	Unclassified proteins (20)	Galactose metabolism (1); Pentose and glucorate interconversion (1),



Future work

- Improve cluster interpretation and validation
- Other similarity assessment methods
- Cluster-based assignment of function to uncharacterized genes
- Other integration (machine learning) methods
- Different model organisms
- Other applications of GO-driven similarity: Co-expression validity assessment, relationship with other functional properties.

Summary

- A method to reconstruct networks using similarity information extracted from the GO and the *Saccharomyces* Genome Database (SGD).
- GOSN represent significant features of real functional networks
- These networks, in combination with other relevant predictive resources, have the potential to improve the overall predictive ability of integrated networks
- Integrated networks comprising GOS relationships contain more meaningful clusters than those ignoring GOS-based evidence.

Acknowledgments

- LHNCBC, NLM, NIH:

Dr. Donald King (Acting Director)

May Cheh and Rob Logan (Program Coordinators)

Research collaborators:

- Olivier Bodenreider, NLM, NIH
- Haiying Wang and Huiru Zheng, UU
- Alban Chesneau, EMBL-Grenoble

Contact, additional information:

fj.azuaje@ieee.org

http://ijsr32.infj.ulst.ac.uk/~e10110731